PointLLM: Empowering Large Language Models to Understand Point Clouds

Abstract

The unprecedented advancements in Large Language Models (LLMs) have shown a profound impact on natural language processing but are yet to fully embrace the realm of 3D understanding. This paper introduces PointLLM, a preliminary effort to fill this gap, enabling LLMs to understand point clouds and offering a new avenue beyond 2D visual data. PointLLM understands colored object point clouds with human instructions and generates contextually appropriate responses, illustrating its grasp of point clouds and common sense. Specifically, it leverages a point cloud encoder with a powerful LLM to effectively fuse geometric, appearance, and linguistic information. We collect a novel dataset comprising 660K simple and 70K complex point-text instruction pairs to enable a two-stage training strategy: aligning latent spaces and subsequently instruction-tuning the unified model. To rigorously evaluate the perceptual and generalization capabilities of PointLLM, we establish two benchmarks: Generative 3D Object Classification and 3D Object Captioning, assessed through three different methods, including human evaluation, GPT-4/ChatGPT evaluation, and traditional metrics. Experimental results reveal PointLLM's superior performance over existing 2D and 3D baselines, with a notable achievement in human-evaluated object captioning tasks where it surpasses human annotators in over 50% of the samples. Codes, datasets, and benchmarks are available at https://github.com/OpenRobotLab/PointLLM .